2020

What Do Interviewers Learn?: Changes in Interview Length and Interviewer Behaviors over the Field Period

Kristen M. Olson

Jolene Smyth

What Do Interviewers Learn?: Changes in Interview Length and Interviewer Behaviors over the Field Period

Kristen Olson and Jolene D. Smyth

University of Nebraska-Lincoln

Contents

Published in *Interviewer Effects from a Total Survey Error Perspective,* ed. Kristen Olson, Jolene D. Smyth, Jennifer Dykema, Allyson L. Holbrook, Frauke Kreuter, and Brady T. West (2020), Boca Raton: CRC Press, pp. 279-291.

1 Introduction

Interviewers are important actors in telephone surveys. By setting the pace for an interview, interviewers communicate the amount of time and cognitive effort respondents should put into their task. It is well-established that interviewers vary widely in the time they spend administering a survey, and that this time changes over the course of the data collection period as interviewers gain experience (Bohme and Stohr 2014; Kirchner and Olson 2017; Loosveldt and Beullens 2013a, 2013b; Olson and Bilgen 2011; Olson and Peytchev 2007). In particular, interviewers get faster as they gain experience over the field period of a survey.

The within-survey effect of experience on interview length is generally attributed to interviewer learning effects. In particular, a learning effect occurs when interviewers learn how to change their behaviors to more quickly administer questions. This can include positive changes in behaviors over the field period such as error-free administration of questions or negative changes such as shortening questions (i.e., non-standardization) or avoiding positive, time-consuming behaviors like probing or verifying answers (e.g., Bohme and Stohr 2014; Kirchner and Olson 2017; Loosveldt and Beullens 2013a, 2013b; Olson and Peytchev 2007). Other hypotheses about why the length of interview changes over the course of the data collection period, including characteristics of the respondents or interviewers or differential respondent motivation correlated with their response propensity, have not explained away the learning effect (e.g., Kirchner and Olson 2017). However, Kirchner and Olson (2017) found that a measure of the interaction between interviewers and respondents—the number of words spoken by the interviewer and by the respondent —partially mediated the interviewer learning effect.

Despite the well-replicated finding that interviewers speed up over the field period, what behaviors change and whether they explain the decrease in interview length over the course of data collection has not been previously examined in published articles. This chapter examines two research questions:

RQ1: What standardized, nonstandardized, and inefficient interviewer behaviors change over the course of the data collection period?

RQ2: Do these behaviors account for changes in interview length over the course of the data collection period?

To answer these questions, we draw on two nationally representative US telephone surveys of adults. Both surveys were audio-recorded and transcribed. Interviewer and respondent behaviors were coded at the conversational turn level, allowing a detailed examination of the changes in interviewer behaviors over the course of the field period. We focus on interviewer behaviors, as the learning hypothesis focuses primarily on changes by the interviewer, although interviewer behaviors inevitably affect respondent behaviors as well.

2 Hypotheses for Behaviors Affected by Interviewer Learning

There are three main hypotheses about what interviewers may "learn" as they conduct interviews over the course of the field period. First, *interviewers may learn to omit or shorten certain standardized interviewer behaviors (i.e., ''good'' behaviors)*. Standardized behaviors include reading questions exactly as worded, using nondirective probes, repeating the respondents' answers to verify what they said, clarifying the question wording, and providing appropriate feedback to the respondent (Fowler and Mangione 1990). The standardized "good" behaviors may be eliminated as interviewers learn what may be shortcut, become bored or frustrated with certain questions, think that certain questions are emotionally draining for follow-up, or think that they remember the question wording, and thus do not read the item on the questionnaire directly (Kaplan and Yu Chapter 5; Ongena and Dijkstra 2007). As the field period progresses and interviewers learn from previous respondents' answers, they also may be more likely to enter a response that is not directly codable rather than probe nondirectively for a codable response (Ongena and Dijkstra 2007). Finally, interviewers may reduce their use of trained techniques that are used less frequently during interviews (e.g., probing), especially experienced interviewers for whom training is more distant (Olson and Bilgen 2011; Tarnai and Moore 2008; van der Zouwen, Dijkstra, and Smit 1991).

Second, *interviewers may learn to become more efficient at administering questions by reducing or eliminating seemingly extraneous behaviors,* including stuttering or disfluencies while reading questions (Olson and Peytchev 2007). Interviewers may also reduce or eliminate extraneous laughter in an effort to shorten their interactions with respondents. They may do so because they place a greater premium on efficiency than rapport, or because their own enthusiasm conducting the survey wears thin over time (Cleary, Mechanic, and Weiss 1981; Houtkoop-Steenstra 1997). For the same reasons, interviewers may reduce their use of verbal pleasantries, personal disclosures, flattery, and digression. Interviewers may also reduce or eliminate task-related feedback (e.g., "let me just get this down") as early bugs in the interview hardware or software are corrected or as they become more efficient in navigating the interview system or entering responses. Task-related feedback may also be reduced if interviewers think it is not helpful for maintaining rapport or guiding respondents through the interview. Notably, these inefficiency-related behaviors may not be part of interviewer training, but happen as part of normal conversation.

Finally, *interviewers may learn to increase the use of nonstandardized, time-saving behaviors* such as changing the question wording (including making major changes or skipping questions), directively probing inadequate answers, changing answers when verifying them, and interrupting respondents. Although interviewers are specifically trained to avoid these behaviors, nonstandardized behaviors are ubiquitous in standardized interviews (Edwards, Sun, and Hubbard Chapter 6; Ongena and Dijkstra 2006). For instance, interviewers may be more likely to adopt practices, such as directively probing an uncodable answer, in order to advance through the interview more quickly (van der Zouwen, et al. 1991).

It is also possible that these behaviors will differ for landline versus cell phone interviews, as previous research has illuminated differences in interviewer and respondent conversational behaviors across these devices (Timbrook, Smyth, and Olson 2018).

3 Data and Methods

This chapter builds on Kirchner and Olson (2017), using the same two telephone surveys. First, the Work and Leisure Today 1 (WLT1) Survey was a land line random digit dial (RDD) telephone survey conducted by AbtSRBI between July 31 and August 28, 2013 (n = 450, AAPOR RR3 = 6.3%). WLT1 contained questions about the respondents' employment, leisure activities, technology use, and demographics. It was deliberately designed to have some highly problematic questions, including difficult and unknown terms, sensitive items, and complex questions. To facilitate model estimation (van Breukelen and Moerbeek 2013), we restricted analyses to the 19 interviewers who conducted at least 10 interviews (n = 435 respondents).

Second, the Work and Leisure Today 2 (WLT2) Survey was a dual frame RDD telephone survey conducted by AbtSRBI during September 2015 (n = 902, landline = 451, AAPOR RR3 = 9.4%; cell phone = 451, AAPOR RR3 = 7.1%). This survey also contained questions about work, leisure, technology use, and demographics, but it did not include many of the highly problematic questions found in WLT1. Although these surveys are called WLT1 and WLT2, the samples are fully independent; that is, there is no longitudinal component. The WLT2 questionnaire contained two versions with alternative experimental questionnaire designs and question wording on many questions. As with WLT1, we restricted the analysis to the 26 interviewers with at least 10 completed interviews (n == 896 respondents). Each of the surveys was audio-recorded, transcribed, and behavior-coded at the conversational turn level using Sequence Viewer (Dijkstra 2016). Eight fields were coded by trained undergraduate coders, with a 10% subsample of interviews in each study coded by two master coders; we use seven codes in this chapter. For each conversational tum, coders identified the actor (e.g., interviewer), the initial action (e.g., question asking), an assessment of that initial action (e.g., question read with changes), details of that action (e.g., changes were major), whether a particular actor laughed, either on its own or as part of a conversational turn, whether there were any disfluencies (including uhs, ums, and stuttering), and whether one actor interrupted the other actor. Kappa values exceeded 0.8 for. most codes; assessments of the initial action exceeded 0.5 (see Online Appendix for details).

3.1 Creating Behavior Measures

Because we are interested in explaining total interview length, we aggregate behaviors to the interview level. There are two approaches to examining summary measures of behaviors at the interview level. First, we can examine the *total number of conversational turns* on which each type of behavior occurred. The number of conversational turns with a given behavior is a measure of how much conversation occurred due to this behavior within a single interview (i.e., an interview-level count). This measure accounts for all behaviors that occurred during the interview (e.g., multiple probing turns on the same question will be counted), but obscures whether the behaviors occurred on only a few questions or on many questions during the interview. The second approach is to use a count of the *total number of questions* on which an individual behavior occurred within a single interview. This question-level count cannot account for multiple occurrences of a behavior within a question, but it provides a measure of whether the behavior occurred on only a few questions or on many questions throughout the interview. The two measures are highly correlated. We use the question-level count in the current analysis. Results are similar for the count of the number of conversational turns across the interview (available on request).

3.2 Dependent Variables

We examine two sets of dependent variables. The first, corresponding to RQ1, are the *interviewer behaviors*. We counted the total number of questions on which each behavior occurred at least once across the entire questionnaire (an average of 50 questions per respondent in WLT1 and 51 questions per respondent in WLT2). Our five measures of standardized "good" behaviors include exact question reading, nondirective probes, exact verification, appropriate clarification, and appropriate feedback. Our five measures of inefficiency behaviors include stuttering and repairs during question reading, disfluencies, "pleasant talk," task-related feedback, and laughter. Finally, we have five measures of nonstandardized behaviors, including (major) changes in question wording, directive probes, inadequate verification (paraphrasing), and interruptions. The operationalization and distribution for each of these behaviors are shown in **Table 1.**

Table 1 Definition and Mean Number of Questions with Each Interviewer Behavior

	Definition		*WLTI*		*WLT2*	
	Action	*Assessment*	*Mean*	*%*	*Mean*	*%*
Standardized behaviors						
Exact question reading	Question asked	Question asked exactly as worded	23.36	46.23	36.99	72.22
Nondirective probes	Probe	Repeat entire question, Repeat part of question, Repeat response options, "Take your best guess," or Ask for explicit response	8.37	16.94	7.73	15.20
Exact verification	Verification	Verifies by repeating respondent's answer exactly	7.92	15.84	7.17	14.01
Appropriate clarification	Clarify	"Whatever it means to you," Provide definition exactly as worded, or Clarifying unit	1.94	3.84	0.45	0.86
Appropriate feedback	Actor = Interviewer + Feedback	Affirmation, Short acknowledgment, or Long motivational feedback	19.52	39.16	22.57	44.14
Inefficiency behaviors						
Stuttering during question reading	Question asked	Read question with stutters, or Read response options with stutters	2.74	5.51	2.42	4.74
Disfluencies	Any disfluencies		13.31	26.77	11.72	22.96
Pleasant talk	Feedback	Personal disclosure, Flattery, or Digression	0.49	0.99	0.65	1.27
Task-related feedback	Feedback	Task-related feedback, Telephone quality, or Time-related feedback	0.88	2.07	1.37	2.77
Laughter	Interviewer laughed + Both laughed		2.32	4.63	3.12	6.05
Nonstandardized behaviors						
Minor changes in question wording	Question asked	Read question with changes - Slight changes	15.40	30.79	4.88	9.60
Major changes in question wording	Question asked	Read question with changes - Major changes	5.44	11.01	6.27	12.15
Directive probes	Probe	Question repeated with changes, or Other directive probes	2.63	5.36	1.14	2.28
Inadequate verification	Verification	Repeated respondent's answer with changes	3.23	6.53	1.83	3.58
Interruptions	Interviewer interrupts respondent		5.63	11.35	3.30	6.49

All behaviors evaluated only on conversational turns during which the interviewer was identified as the actor.

The second dependent variable is *interview length in minutes.* To address outliers (Yan and Olson 2013), interview length was trimmed at the 1st and 99th percentiles. The mean interview length was 12.65 minutes for WLT1 and 13.36 minutes for WLT2.

3.3 Primary Independent Variable: Within-Survey Experience

Within-survey experience is the primary measure of whether interviewers learn over the course of the field period. Because we expect that the effect of learning will be larger at the beginning of the field period than at the end of the field period (Olson and Peytchev 2007), we include a log-transformed ordinal counter for interview order (i.e., 1 for the first interview for an interviewer, 2 for the second, etc.). This counter ranges from 1 to 27 in WLT1 and from 1 to 79 in WLT2.

3.4 Control Variables

Because respondent characteristics and response propensity also differ over the field period and across the two studies, we include the following control variables: an overall measure of interviewer experience (i.e., less than one year versus one year or more), the interviewer-level cooperation rate, other interviewer characteristics (race, gender, worked primarily weekday evening shifts), respondent characteristics (sex, age, education, employment status, income, household size, parental status, volunteer status, computer usage), and measures of response propensity (item nonresponse rate, whether the household ever refused, whether the interview was completed at first contact, number of call attempts, time of day the interview was completed). Finally, the number of answers that were changed by the interviewer as recorded in the paradata are included as control variables for both studies. In WLT2, we also included indicators for which experimental questionnaire was used and whether the interview was conducted on a land line or a cell phone.

3.5 Analytic Strategy

We estimate hierarchical two-level random intercept models accounting for the clustering of respondents within interviewers (e.g.,

Raudenbush and Bryk 2002). For the interviewer behaviors, we estimate two-level Poisson models with a log link and a random intercept due to interviewers, with the number of questions asked to each respondent as the exposure variable (see the online supplementary materials). These models are estimated using the **`mepoisson`** procedure in Stata 15.1. For interview length, we estimate a two-level linear model using the mixed procedure in Stata 15.1 with a random intercept due to interviewers (see the online supplementary materials). In these models, the interview behaviors are grand-mean centered. They are initially included as separate groups (standardized, inefficiency, nonstandardized) and then combined into a single model.

4 Results

4.1 RQ1: What Interviewer Behaviors Change over the Course of the Data Collection Period?

We start by addressing RQ1. We focus only on the interview order (within-survey experience) coefficient in our discussion below. The full models are in the online supplementary materials. **Table 2** contains the coefficients from the log(interview order) term in both WLT1 and WLT2.

We start with standardized interviewing behaviors. We see notable differences across the two surveys. In WLT1, there is no change in the number of questions on which standardized interviewer behaviors occur across the data collection period at traditional $p<.05$ levels. In WLT2, on the other hand, there are statistically significant decreases in the number of questions on which nondirective probes, exact verification, and appropriate feedback occur as interviewers gain within-study experience. The difference in coefficients between WLT1 and WLT2 is statistically significant for nondirective probes ($z = 2.23$, $P = 0.026$) and exact verification ($z = 2.55$, $P = 0.011$). To understand the magnitude of these changes, we examine predicted marginal effects. The average workload among interviewers in WLT2 who conducted at least ten interviews was 34.5 interviews. As such, we examine changes from the 1st to the 30th interview. On average, the predicted number of questions in which interviewers use nondirective probes decreases

Table 2 Unstandardized Coefficients from Log(Interview Order) Predicting Count of Questions with Interviewing Behaviors

	WLTI		*WLT2*		*z-Value for Test Across Surveys*
	Coef.	*SE*	*Coef.*	*SE*	
Standardized interviewing behaviors					
Exact question reading	0.017	0.014	0.001	0.007	1.08
Nondirective probes	0.020	0.023	-0.033*	0.015	2.23*
Exact verification	0.020	0.026	-0.051**	0.016	2.55*
Appropriate clarification	0.091	0.050	-0.034	0.061	1.77
Appropriate feedback	-0.010	0.015	-0.035****	0.009	1.33
Inefficiency behaviors					
Stuttering during question reading	-0.201****	0.039	-0.201****	0.025	0.10
Disfluencies	-0.062**	0.018	-0.058****	0.012	0.01
Pleasant talk	-0.086	0.096	-0.098	0.051	-0.11
Task-related feedback	-0.151*	0.061	-0.052	0.035	-1.30
Laughter	-0.162****	0.043	-0.084****	0.024	-1.70
Nonstandardized behaviors					
Minor changes in question reading	0.032	0.018	0.050*	0.020	-0.98
Major changes in question reading	-0.050	0.028	0.036*	0.017	-2.49*
Directive probes	0.100*	0.042	-0.087*	0.039	4.04****
Inadequate verification	-0.176****	0.036	-0.106****	0.030	-1.28
Interruptions	-0.046	0.028	-0.069**	0.023	0.99

* p < .05 ; ** p < .01 ; *** p < .001 ; **** p < .0001

11% from 8.2 in the 1st interview to 7.3 in the 30th interview. The use of exact verification decreases 16% from a predicted average of 9.2 questions in the 1st interview to 7.7 in the 30th interview. Appropriate feedback decreases about 11% from being used on an average of 24.4 questions in the 1st interview to 21.6 in the 30th interview.

We now tum to inefficiency behaviors, shown in the middle of Table 2. In both WLT1 and WLT2, there are fewer questions with inefficiency behaviors as interviewers gain within-study experience. The coefficients do not statistically significantly differ across the two studies. In both studies, just over 4 questions are read with stutters on the 1st interview, compared to about 2.1 questions with stutters by the 30th interview. Although the number of questions on which a disfluency occurs differs for WLT1 and WLT2, the rate of decline is similar, 18-19%, across the field period in both studies - falling from a predicted average of 15.8 questions with some sort of disfluency on the 1st interview to 12.8 questions with disfluencies by the 30th interview in WLT1 (13.9 to 11.4 in WLT2). The number of questions with laughter

also declines across the field period, from 3.9 (4.4, WLT2) questions on the 1st interview to 2.2 (3.3, WLT2) questions on the 30th interview in WLT1. The rate of task-related feedback declines from 1.5 questions on the 1st interview to 0.88 questions on the 30th interview in WLT1 but not WLT2. There is no statistical change in pleasant talk across the field period in either study.

Finally, we look at nonstandardized interviewing behaviors, shown at the bottom of Table 2. Across the two surveys, there are mixed changes in nonstandardized behaviors. In both studies, the rate of inadequate verification behaviors declines across the field period, from an average of 4.8 questions on the 1st interview in WLT1 (2.4, WLT2) to an average of 2.6 questions by the 30th interview (1.6, WLT2). The rate of interruptions also declines by about one question over the field period in both studies (WLT1: 6.4 questions to 5.4 questions; WLT2: 4.0 questions to 3.1 questions). In both studies, the number of questions with minor changes in question wording increases by about 1.5 questions over the field period (from 14.5 to 16.2 in WLT1 and from 4.3 to 5.1 in WLT2). None of the interview order coefficients differ between the two studies for these outcome variables.

There appears to be a trade-off between major changes in question reading and in directive probes in the two studies, with statistically significant differences in the interview order coefficients between the two studies. In WLT1, major changes in question reading decline by about 1 question over the field period (from 6.3 questions at the 1st interview to 5.3 questions at the 30th interview), whereas the use of directive probes increases by about 1 question (from 2.1 questions at the 1st interview to 3.0 questions at the 30th interview). In WLT2, the pattern is the opposite - major question reading changes increase (from 5.6 questions to 6.3 questions) and directive probes decrease (from 1.4 to 1.1 questions) (z-test for the difference between interview order coefficients in WLT1 and WLT2: directive probes: $z = 4.04$, $p < .0001$; major changes: $z = -2.49$, $p = 0.013$).

In sum, interviewers do change behaviors as they gain experience over the field period. They become more efficient in administering questions, having fewer questions with stutters, disfluencies, and laughter. Interviewer experience over the field period also changes both standardized and nonstandardized behaviors. In both studies, we see increases in minor changes in question wording and

decreases in interruptions and use of inadequate verification behaviors. There is not a consistent increase in the use of adequate verification behaviors - rather, these behaviors go away. Other changes in nonstandardized behaviors are less consistent across the two studies, with a trade-off between major changes in question wording and directive probes.

4.2 RQ2: Do Interviewer Behaviors Account for Changes in Survey Length over the Course of the Data Collection Period?

We now turn to the question of whether the observed changes in interviewer behaviors explain the changes in survey length over the course of the data collection period. To answer this question, we examine whether the interview order coefficient predicting survey length changes in magnitude as groups of behaviors are included in the model (Aneshensel 2013, p. 184; mediation models for each behavior individually are included in Online Appendix 20C). As seen in **Table 3,** interviewer behaviors only partially explain the change in interview length over the course of the field period. Each group of behaviors reduces the interview order coefficient by about 14-50%. The largest reduction in the coefficient for interview order comes with the inclusion of the inefficiency behaviors in WLT1, reducing the interview order coefficient by 52.4%. However, when all of the interviewer behaviors are included in the same model, this same magnitude reduction in the interview order coefficient is not observed, especially in WLT1. In WLT1, inclusion of the standardized behaviors *increases* the learning effect on length of interview, whereas the other behaviors *explain* the learning effect. Thus, the combined effects "cancel out" in the overall model. In sum, these 15 interviewer behaviors partially mediate the learning effect, but do not completely account for changes in the length of interview over the course of the field period.

4.3 Variance Components

There is significant variation across interviewers and respondents in the length of the interview. As shown in **Table 4**, the interviewer behaviors examined here explain between 21% and 32% of the variance

Table 3 Log(Interview Order) Coefficients Predicting Length of Interview with Interviewing Behaviors and Percent Change from Model with No Behaviors

	WLT1		*WLT2*	
	Log(Interview Order) Coefficient	*% Reduction from No Behaviors*	*Log(Interview Order) Coefficient*	*% Reduction from No Behaviors*
No behaviors, no controls	-0.189		-0.537****	
No behaviors, with controls	-0.443**		-0.855****	
Including only standardized behaviors	-0.627****	-41.5	-0.688****	19.5
Including only inefficiency behaviors	-0.211	52.4	-0.706****	17.4
Including only nonstandardized behaviors	-0.382*	13.8	-0.733****	14.3
Including all behaviors	-0.441**	0.5	-0.607****	29.0

+ p < .10 ; * p < .05 ; ** p < .01 ; *** p < .001 ; **** p < .0001

Negative percent reductions indicate an increase in the coefficient.

Table 4 Variance Components for Interview Length Models

	WLT1				*WLT2*			
	Interviewer		*Respondent*		*Interviewer*		*Respondent*	
	Var. comp.	*% change*	*Var. comp.*	*% change*	*Var. comp.*	*% change*	*Var. comp.*	*% change*
No behaviors	1.828		8.195		2.009		6.166	
Standardized behaviors only	2.847	56	4.463	-46	2.733	36	4.429	-28
Inefficiency behaviors only	2.980	63	6.480	-21	1.841	-8	5.074	-18
Nonstandardized behaviors only	0.899	-51	5.186	-37	1.597	-21	4.383	-29
All behaviors	1.237	-32	3.768	-54	1.594	-21	3.555	-42

in interview length at the interviewer level and between 42% and 54% of the variance in interview length at the respondent level. The inclusion of standardized behaviors alone actually increases the variance at the interviewer level in both studies, as does the inclusion of only inefficiency behaviors in WLT1. Nonstandardized behaviors explain the most variation in length across interviewers in both studies and across respondents in WLT2.

5 Conclusion

This chapter aimed to answer two questions - do interviewer behaviors change over the field period, and do the changes in these interviewer behaviors account for the learning effect observed in the shortening of the interview length over the field period? We found clear confirmation for the first question - interviewer behaviors do change over the course of the data collection period. We also found, reassuringly, that interviewer behaviors are related to interview length. However, interviewer behaviors do not fully explain the learning effect.

First, interviewers do not consistently lose standardized behaviors over the field period across these studies. This is good news. Where there are losses in standardized behaviors in WLT2, it appears to be in feedback behaviors (e.g., ok; thank you), as well as some minor decreases in nondirective probing and verification behaviors. These changes in standardized behaviors explain between none and about 20% of the change in interview length. Second, interviewers do become more efficient in administering surveys over the field period. These changes in inefficiency behaviors explain 17-44% of the change in interview length over the field period. Notably, the inefficiency behaviors alone render the interview order coefficient non-significant at traditional $p < .05$ levels. Finally, interviewers do change in their use of nonstandardized behaviors. There is evidence of an increase in minor changes of question wording over the field period in both studies, perhaps because interviewers are further away from training. Alternatively, interviewers may be learning that respondents have problems with certain questions and preemptively changing the question wording to anticipate where those problems occur. Other nonstandardized behaviors such as inadequate verification and interruptions decrease over the field period. We also see potential trade-offs between major changes in question wording and directive probes in these surveys. Collectively, nonstandardized behaviors explain about 14% of the change in interview length over the field period.

How interviewer behaviors are related to interview length is more complicated than simply the number of questions on which these behaviors occur. Other factors that we have not yet examined

are whether interviewers become faster at selecting among the various behaviors they use after question asking or whether they are completing the behaviors themselves more quickly (e.g., do they probe or clarify with fewer words? with faster paced speech?). We also did not examine how question characteristics themselves affect the occurrence of these behaviors. There are clear differences in the prevalence of certain interviewer behaviors across WLT1 and WLT2. Although the content of the questionnaires was similar over these two studies, the questionnaires varied in difficult terms, sensitive questions, and other question characteristics. It may be that inefficiency behaviors are largely properties of individuals' conversational norms and basic linguistic practices, whereas standardized and nonstandardized behaviors are more sensitive to properties of the questions themselves. There remains much future research to do in this area.

We note several limitations with this study. First, we looked only at changes in interviewer behaviors, but many interviewer behaviors occur in reaction to respondent behaviors, either because of the requirements of standardized interviewing (e.g., probing to obtain an answer) or to maintain rapport (Garbarski, Schaeffer, and Dykema 2016). That is, inferential problems may arise because of how the behaviors themselves unfold during an interview, where one behavior may be a trigger for another behavior. As such, future research should examine changes in respondent behaviors as well. Second, although the results largely replicate when we aggregate behaviors to the level of number of conversational turns rather than the number of questions on which the behavior occurs, the number of conversational turns may be a better reflection of the length of the interview. Third, there is sensitivity in our conclusions depending on the collection of behaviors that are included in these models. Some of this is due to potential overlap of constructs represented by the various behaviors (e.g., different types of question-asking behaviors), creating issues of multicollinearity if the number of questions asked in the survey is also included in the duration models. Despite these limitations, a significant strength of this study is that it pursued the goal of replication by examining two surveys conducted two years apart with different interviewing teams. Also, this study used behavior codes to directly evaluate what is happening in the survey interview itself, an

incredibly time-consuming and expensive method to produce, for both surveys. However, both surveys were conducted by the same organization. Future research will examine surveys conducted by a different organization.

Even with these limitations, our findings do yield practical implications. Most notably, interviewing efficiency may be gained by changing interviewer training practices. Currently, survey-specific interviewer trainings often involve round robins where interviewers read aloud a single question, but may not read through the entire questionnaire more than once or twice before the start of the field period (Tarnai and Moore 2008). These round robins may not give interviewers enough question-specific practice to reduce stuttering and disfluencies prior to live interviewing. Requiring interviewers to complete entire practice interviews multiple times could help eliminate some inefficiencies prior to live interviewing, thus ensuring more efficient delivery even on early interviews and mitigating the change in inefficient behaviors over the field period. Some organizations also retrain interviewers during the field period. This study suggests retraining on nonstandardized and inefficient behaviors could further reduce the length of the interview. .

This study is the first to evaluate how a wide range of interviewer behaviors change over the course of a field period, both individually and as related to interview length. We find that interviewers do change behaviors and that these behaviors partially explain changes in interview length over the field period. We see notable decreases in a wide variety of interviewer behaviors over the course of the data collection period. These decreases suggest less interaction overall between an interviewer and a respondent later in the field period. However, that these 15 theoretically derived interviewer behaviors do not fully explain changes in the length of the interview suggests that there is more about the interaction that is important for future research.

Acknowledgments This work was supported by the National Science Foundation Grant No. SES-1132015. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Thanks to Antje Kirchner, Beth Cochran, Jinyoung

Lee, Jerry Timbrook, Amanda Ganshert, and Alexis Swendener for research assistance. Thanks to all of our transcriptionists and behavior coders for their amazing work. Previous versions of this chapter were presented at the 2015 AAPOR conference and the 2018 Joint Statistical Meetings.

References

Aneshensel, C. S. 2013. *Theory-Based Data Analysis for the Social Sciences.* Thousand Oaks, CA: Sage Publications.

Bohme, M., and T. Stohr. 2014. Household interview duration analysis in CAPI survey management. *Field Methods 26(4):390--405.*

Cleary, P. D., D. Mechanic, and N. Weiss. 1981. The effect of interviewer characteristics on responses to a mental health interview. *Journal of Health and Social Behavior 22(2):183-193.*

Dijkstra, W. 2016. Sequence Viewer.

Fowler, F. J., and T. W. Mangione. 1990. *Standardized Survey Interviewing: Minimizing Interviewer-Related Error.* Newbury Park, CA: Sage Publications.

Garbarski, D., N. C. Schaeffer, and J. Dykema. 2016. Interviewing practices, conversational practices, and rapport: Responsiveness and engagement in the standardized survey interview. *Sociological Methodology 46(1):1-38.*

Houtkoop-Steenstra, H. 1997. Being friendly in survey interviews. *Journal of Pragmatics 28(5):591-623.*

Kirchner, A., and K. Olson. 2017. Examining changes of interview length over the course of the field period. *Journal of Survey Statistics and Methodology 5:84--108.*

Loosveldt, G., and K Beullens. 2013a. 'How long will it take?' An analysis of interview length in the fifth round of the European Social Survey. *Survey Research Methods 7(2):69-78.*

Loosveldt, G., and K Beullens. 2013b. The impact of respondents and interviewers on interview speed in face-to-face interviews. *Social Science Research 42(6):1422-1430.*

Olson, K, and I. Bilgen. 2011. The role of interviewer experience on acquiescence. *Public Opinion Quarterly 75(1):99-114.*

Olson, K, and A. Peytchev. 2007. Effect of interviewer experience on interview pace and interviewer attitudes. *Public Opinion Quarterly 71(2):273-286.*

Ongena, Y. P., and W. Dijkstra. 2006. Methods of behavior coding of survey interviews. *Journal of Official Statistics 22:419-451.*

Ongena, Y. P., and W. Dijkstra. 2007. A model of cognitive processes and conversational principles in survey interview interaction. *Applied Cognitive Psychology* 21 (2):145-163.

Raudenbush, S. W., and A. S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA: Sage.

Tarnai, J., and D. L. Moore. 2008. Measuring and improving telephone interviewer performance and productivity. In: *Advances in Telephone Survey Methodology,* ed. J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. de Leeuw, L. Japec, P. J. Lavrakas, M. W. Link, and R. L. Sangster, 359-384. Hoboken, NJ: John Wiley & Sons, Inc.

Timbrook, J., J. Smyth, and K Olson. 2018. Why do mobile interviews take longer? A behavior coding perspective. *Public Opinion Quarterly 82(3):553-582.*

van Breukelen, G., and M. Moerbeek. 2013. Design considerations in multilevel studies. In: *The SAGE Handbook of Multilevel Modeling,* ed. M. A. Scott, J. S. Simonoff, and B. D. Marx, Chapter 11. London: SAGE Publications Ltd.

van der Zouwen, J., W. Dijkstra, and J. H. Smit. 1991. Studying respondent-interviewer interaction: The relationship between interviewing style, interviewer behavior, and response behavior. In: *Measurement Errors in Surveys,* ed. P. Biemer, R. M. Groves, L. Lyberg, N. A. Mathiowetz, and S. Sudman, 419-438. New York: John Wiley & Sons, Inc.

Yan, T., and K. Olson. 2013. Analyzing paradata to investigate measurement error. In: *Improving Surveys with Paradata: Analytic Uses of Process Information,* ed. F. Kreuter, 73-95. Hoboken, NJ: John Wiley & Sons, Inc.